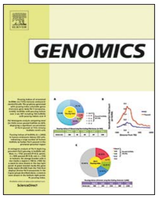
Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno



Selection shapes the patterns of codon usage in three closely related species of genus *Misgurnus*

Shaokui Yi, Yanhe Li, Weimin Wang*

College of Fisheries, Key Lab of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education/Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Huazhong Agricultural University, Wuhan 430070, China

ARTICLE INFO

Keywords:
Misgurnus
Codon usage
Codon pair
Selection

ABSTRACT

Neutrality plots revealed that selection probably dominates codon bias, whereas mutation plays only a minor role, in shaping the codon bias in three loaches, *Misgurnus anguillicaudatus*, *M. mohoity*, and *M. bipartitus*. These three species also clearly showed similar tendencies in the preferential usage of codons. Nineteen, nine, and 14 preferred codon pairs and 179, 182, and 173 avoided codon pairs were also detected in *M. anguillicaudatus*, *M. bipartitus*, and *M. mohoity*, respectively, and the most frequently avoided type of cP3-cA1 dinucleotide in these species was nnUAnn. The expression-linked patterns of codon usage revealed that higher expression was associated with higher GC_3, lower ENC, and a smaller proportion of amino acids with high size/complexity (S/C) scores in these three species. These results elucidate selectively driven codon bias in *Misgurnus* species, and reveal the potential importance of expression-mediated selection in shaping the genome evolution of fish.

1. Introduction

Three species of loach, *Misgurnus anguillicaudatus*, *M. mohoity*, and *M. bipartitus*, are distributed throughout China, except on the Tibetan Plateau [1,2]. *Misgurnus anguillicaudatus* is widely distributed in the middle and lower reaches of the Yangtze River Basin, whereas *M. bipartitus* is restricted to north of the Yellow River in China, and *M. mohoity* only occurs in the Amur River Basin in northeast China, Mongolia, and the Far East region of Russia [3–5]. Given their high nutritional value and delicacy, these loach species are widely consumed in local markets. Therefore, they are considered to be three of the most valuable fisheries resources and have become potentially important targets for aquaculture, domestication, and stock enhancement. Despite their economic and ecological importance, studies of these closely related species remain limited. Previous studies primarily focused on the molecular markers [6–9] (e.g., microsatellites, isozymes, and mitochondrial genes), polyploid breeding [10–12], and gene functions [13–15] of these species. However, much remains unknown about these three loach species, and a broader understanding of the evolutionary status of the *Misgurnus* genomes is required. Since the development of the next-generation sequencing technology, RNA-seq has been widely used as an efficient and accessible approach to investigating the evolutionary dynamics of protein-coding genes when no genomic information is available.

Triplet codons are the basic coding units of mRNA. Because the genetic code is degenerate, a protein sequence can be encoded by many different synonymous mRNA sequences. Synonymous codon usage was once thought to be functionally neutral, but evidence now indicates that synonymous codons are not used randomly in translation [16,17]. Previous studies [18–20] proposed that bias in synonymous codon usage can arise from mutational pressure or selective forces that favor efficient and accurate translation. In particular, codon usage bias correlates with transfer RNA abundance, gene copy numbers, and gene expression levels [17,21,22]. Selection favors specific codons that promote the efficient and accurate translation of genes that are expressed at high levels. In this context, the presence or absence of optimal codons can indicate whether selection for translational efficiency and accuracy plays a major role in an organism's genomic evolution. Many previous studies of codon usage have documented codon usage in a wide variety of model organisms, including *Caenorhabditis* [19,23], *Drosophila* [24,25], and *Arabidopsis* [18], but the studies of codon usage in fish are few and limited [26–28]. Large numbers of sequences can be obtained with the RNA-seq method, providing an opportunity to analyze the codon usage patterns in non-model organisms. Overall, the study of codon usage patterns is crucial for understanding the genomic architecture and molecular evolutionary features of species. In this study, we examined optimal codon and amino acid usage based on the recently available large-scale RNA-seq data for three *Misgurnus* species. Using these data, we investigated the codon usage patterns and whether codon usage and amino acid frequencies have been optimized in the

* Corresponding author.
E-mail addresses: liyanhe@mail.hzau.edu.cn (Y. Li), wangwm@mail.hzau.edu.cn (W. Wang).

http://dx.doi.org/10.1016/j.ygeno.2017.09.004
Received 20 June 2017; Received in revised form 29 August 2017; Accepted 7 September 2017

Please cite this article as: Yi, S., Genomics (2017), http://dx.doi.org/10.1016/j.ygeno.2017.09.004

highly expressed genes of these three closely related organisms.

2. Methods

2.1. Filtering and mining transcriptome data

The RNA-seq data were derived from the transcriptomes of the three loach species, which were sequenced in pooled libraries from six tissues (dorsal muscle, skin, liver, spleen, intestine and brain) of each species, and represented large proportions of the genes in their genomes. Total RNA was isolated from tissues using RNAiso Plus Reagent (TaKaRa, China) according to the manufacturer's protocol. The raw data were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession numbers SRR3744972 (*M. anguillicaudatus*), SRR3744973 (*M. bipartitus*), SRR3744974 (*M. mohoity*). The transcripts were assembled with Trinity [29]. The transcripts without isoforms were identified with BLASTN (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/). The longest sequence from the isoforms and the transcripts without isoforms were chosen for subsequent analysis. We then extracted the full-length coding sequences exceeding 100 codons, with an ATG start codon, a stop codon, and no unknown or ambiguous nucleotides, using EMBOSS [30] and in-house Perl scripts. The CDSs (Coding sequences) were annotated using BLASTX with the non-redundant (NR), Swiss-Prot, and Eukaryotic Orthologous Groups (KOG) proteins databases, and the unannotated sequences were filtered out. The final CDS dataset for each species was used to quantify the expression levels. We mapped reads to CDSs, and measured their expression levels as reads per kilobase million (RPKM) with the RESM program [31].

2.2. Indices of codon usage

The values for GC_i were determined with CAIcal [32]. CDSs from all available fish species in Ensembl (www.ensembl.org/) database were extracted using the data mining tool EnsMart, and the GC contents were calculated by an in-house Perl script. Effective number of codons (ENC), which ranges from 20 (only one synonymous codon is used in the corresponding amino acid) to 61 (when all synonymous codons are used equally), is a measure of the overall codon-use bias of a gene [33]. The relative synonymous codon usage (RSCU) is an index that reflects the overall synonymous codon usage variations among genes [34]. ENC and RSCU were calculated with CodonW (http://codonw.sourceforge.net/). The synonymous codon usage order (SCUO) is based on “Shannon information theory” [35] and varies from 0 to 1; it was calculated with CodonO [36] as 1 minus the ratio of the expected to the observed entropy. The codons with RSCU > 1.5 or relative frequency $> 60\%$ of the frequency of the synonymous codon for the corresponding amino acid are defined as high-frequency codons.

2.3. Identification of high-frequency, preferred, and avoided codon pairs

Codon pair frequencies were calculated from data generated by the Perl scripts proposed by Buchan et al. [37]. The observed frequency of a codon pair is the ratio of the occurrence of a certain pair to occurrence of all 3721 (61×61) codon pairs, and the expected frequency of the 3721 codon pairs is the product of the corresponding expected frequency of each codon. The codons with a relative synonymous codon pair usage (RSCPU) > 1.5 or a relative frequency $> 60\%$ of the frequency of synonymous codon pairs for the corresponding amino acid pairs were identified as high-frequency codon pairs. The RSCPU was calculated by the formula described in Duan et al. [26]. The *P* value was calculated following the formula described in a previous study using the Perl script [38]. The observed (o) and normalized expected (e_{nor}) codon pair counts were converted into residual scores for each codon pair ($[o - e_{nor}] / e_{nor}$) [37]. Hierarchical clustering was used to group individual 5′ (P-site) and 3′ (A-site) codons from every codon pair according to the patterns of similar codon pairing preference based on the residual scores. The ratio of observed frequency to expected frequency (log2), with a ratio cutoff of ± 1 (two-fold changes), was the standard used to identify preferred or avoided codon pairs. The MeV program [39] (http://www.tm4.org/mev.html) was used to detect patterns of codon pair preference.

2.4. Expression-linked patterns of codon usage and amino acid frequencies

To identify the optimal codons for each of the 18 amino acids with synonymous codons, we compared the codon usage for the 5% of genes with the highest and lowest expression. The method of comparison of codon usage patterns among the most highly and lowly expressed CDSs has been widely used in previous studies to investigate the preferences in codon usage that are linked to transcription [40–42]. We calculated the RSCU values for each CDS per expression class. Optimal codons were defined as those with a statistically significant and positive $\Delta RSCU = RSCU_{\text{mean highly expressed CDSs}} - RSCU_{\text{mean lowly expressed CDSs}}$ [40,43,44]. Spearman's correction analysis was performed with R (www.r-project.org/). To assess amino acid preferences, we similarly identified the optimal codon list for each amino acid in the three *Misgurnus* species using RSCU relative to gene expression. We evaluated the biochemical cost of protein synthesis using the Dufton methodology [45], where each amino acid is assigned a size/complexity score (S/C) based on its molecular weight and complexity. The proportion of amino acids with a size/complexity score > 40 (Met, Arg, Cys, His, Phe, Trp, and Tyr) was calculated using Microsoft Excel 2010. The difference in each high S/C score amino acid between the highly and lowly expressed CDSs was evaluated using *t*-test.

3. Results

3.1. Indices of codon usage

In total, 88,842,922, 116,147,180, and 76,502,750 raw reads were generated for *M. anguillicaudatus*, *M. bipartitus*, and *M. mohoity*, respectively. The assembly results are summarized in detail in Table S1. Codon usage analysis was based on > 6000 protein-coding sequences, including 2371 CDSs in *M. anguillicaudatus*, 2382 CDSs in *M. bipartitus*, and 2202 CDSs in *M. mohoity*, after the transcripts of three loaches were filtered. The overall GC contents of these *Misgurnus* species were 0.5164–0.5208, and the GC contents varied among different codon positions, with the highest in GC_3, with values of 0.5441–0.5486, were lowest in GC_2, with values of 0.4965–0.5062, and were intermediate in GC_1, with values of 0.5053–0.5087. The GC_3 content varied across the genes of the three loaches, and their distributions were predominantly unimodal (Fig. 1A). Neutrality plots revealed the relationships between GC_{12} (the average of GC_1 and GC_2) and GC_3 in the three loaches, which may reflect the mutation-selection equilibrium that shapes codon usage. There was a significant negative correlation between GC_{12} and GC_3 in each species ($P < 0.01$; Fig. S1). The slope of the regression line for all CDSs ranged from 0.0857 to 0.1388. These results indicate that the mutation pressure on each loach is not the main force, and that codon bias is affected by natural selection. We then filtered the CDSs with high GC_3 according to the GC_3 values with a cutoff of 60%. To better understand whether GC_3 affects the gene properties in the three loaches, the CDSs with high GC_3, including 584 genes in *M. anguillicaudatus*, 638 genes in *M. bipartitus*, and 609 genes in *M. mohoity*, were classified according to gene ontology (GO). The genes with high GC_3 in three loaches exhibited similar functional categories (Fig. 1B). The category with the highest gene number among three loaches was observed in “binding” category, and the “cellular process” and “metabolic process” categories also contained many CDSs with high GC_3 in all three loaches. SCUO showed a strong “U” nonlinear correlation with GC_3 (Fig. 1C). ENC were calculated for all genes in each species, and showed a strong negative correlation between GC3s and ENC in *M.*